# The JANES project: Tools and resources for linguistic analysis and automatic processing of user-generated content in Slovene

Darja Fišer
University of Ljubljana

Zurich, 17 December 2015

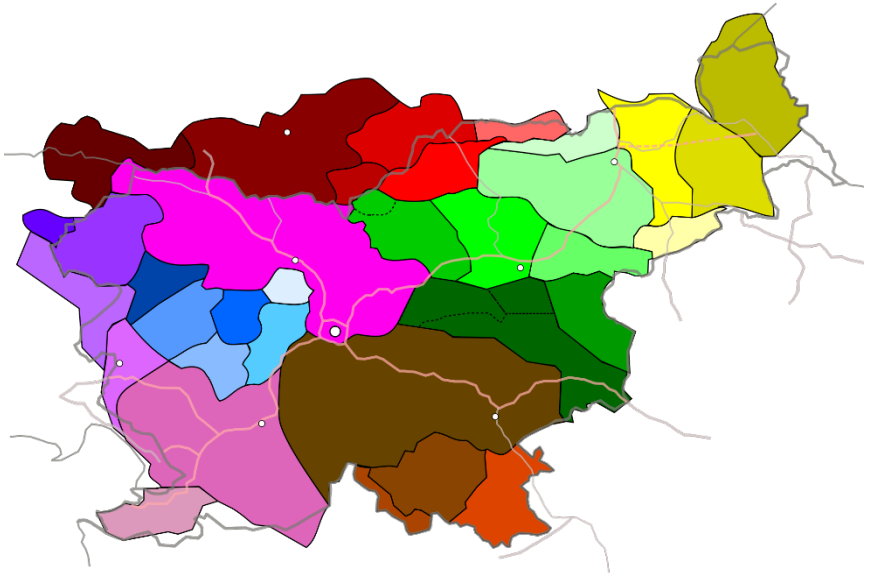
# Outline

1. Project background and goals
2. Construction & annotation of the Janes corpus
3. Construction & annotation of the subcorpus of tweets
4. Work in progress & future work
5. Project activities

# 1. Project background and goals

# Background

- Basic facts
  - 2 mio inhabitants
  - Mediterranean, Alpine, Pannonian and Dinaric regions
  - 7 dialectal groups, 42 dialects
  - prescriptive, normativist culture
- Language resources
  - plenty of corpora (CLARIN.SI & CC)
  - standard language
- The Janes project
  - national basic research project
  - 3 yr, 2014-2017
  - 2 institutions, 8 team members
  - development of resources, tools and methods for the analysis of UGC
  - http://nl.ijs.si/janes



# Research goals

- WP1: Corpus construction

- WP2: Linguistic analysis
  - Task 1: vs. standard Slovene
    - orthography (punctuation, capitalization, spelling)
    - regional variation (PhD)
    - syntax
  - Task 2: vs. spoken Slovene
    - discourse markers
    - interactive elements
  - Task 3: collocations
    - new collocations of old words
    - collocations of new words
  - Task 4: terminology
    - specialisation level & density
    - nonstandard elements in terminology
  - Task 5: semantic shifts
    - sense expansion/narrowing/shifting
    - amelioration/pejoration
  - Task 6: offensive language
    - refugee crisis
    - gay marriage

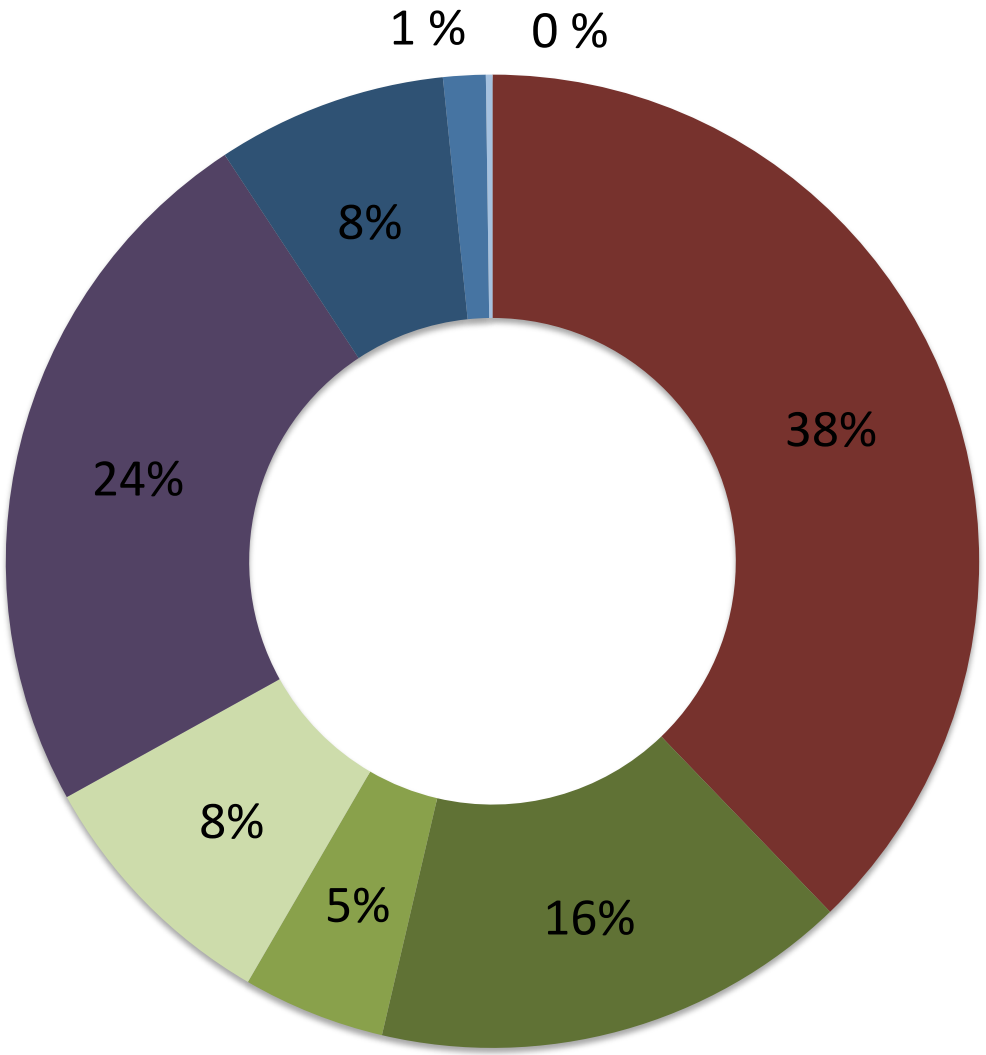
- WP3: NLP tools

# 2. Construction & annotation of the Janes corpus

# Corpus JANES 0.3

- Tweets
  - TweetCat (Ljubešić et al. 2014)
  - Slovene-specific seed words -> Slovene users -> their network
  - metadata: username, time stamp, no. of retweets & favourites
- Forum messages
  - 3 forums: med.over.net, avtomobilizem.com, kvarkadabra.net
  - customized extractors
  - metadata: topic, post URL, post time stamp, username, post id
- News comments
  - 3 news portals: RTV Slo, Mladina, Reporter
  - customized extractors
  - metadata: article url, article id, username, post time stamp, post id
- [ Blogs ]
  - slWaC 2.0 (Erjavec and Ljubešić 2014)
  - “blog” in domain name
  - no metadata, mixed blog text and comments

# Corpus composition by text source

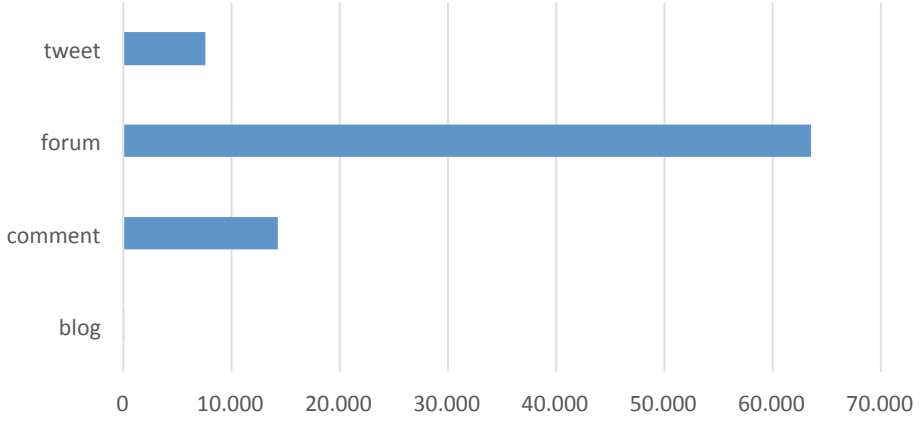



Total: 161 M tokens

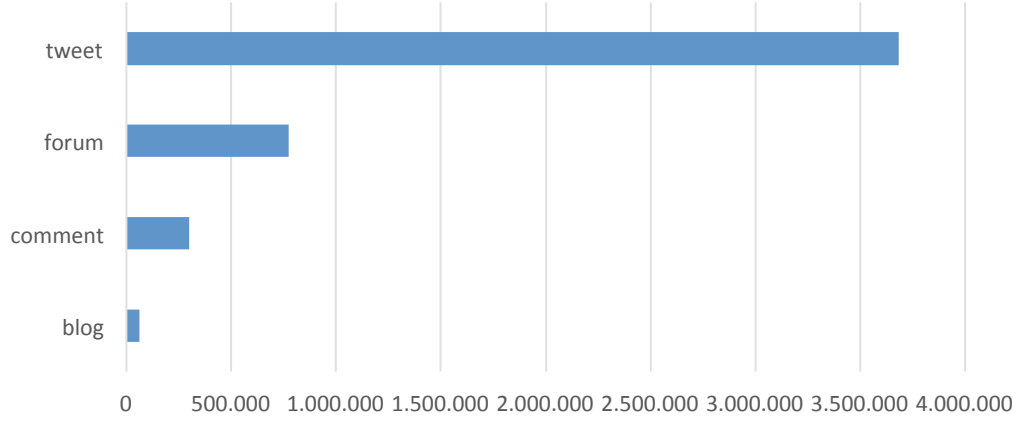
| | |
|---|---|
| Twitter | 61 MT |
| Forumi: Avtomobilizem | 47 MT |
| Forumi: Kvarkadabra | |
| Forumi: Medover.net | |
| Blogi | 38 MT |
| Komentarji: RTV Slo | 15 MT |
| Komentarji: Mladina | |
| Komentarji: Reporter | |

# Corpus composition by authors, texts, words & tokens

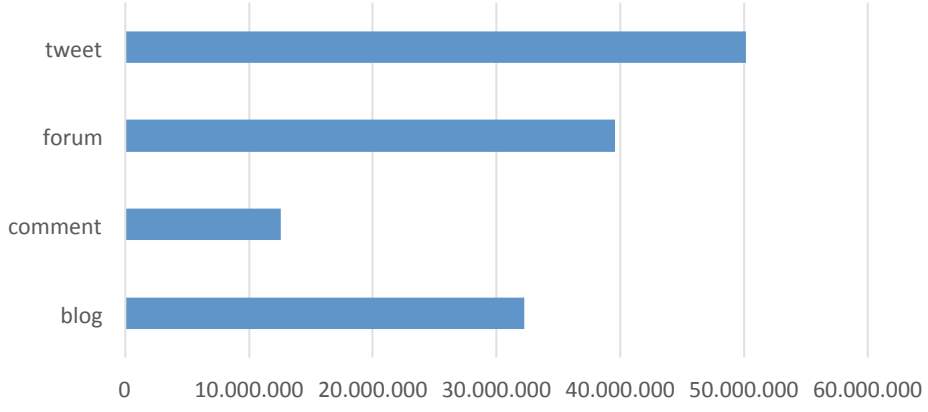
## 85,500 authors



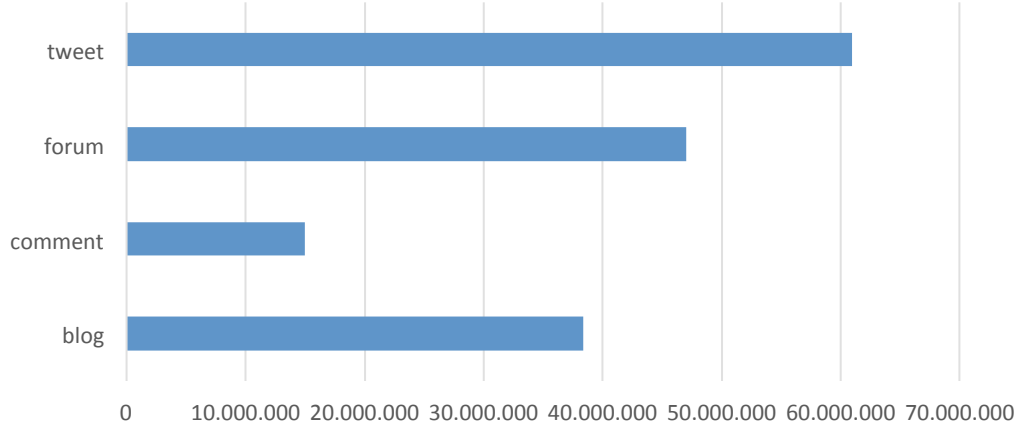

## 4,8M texts



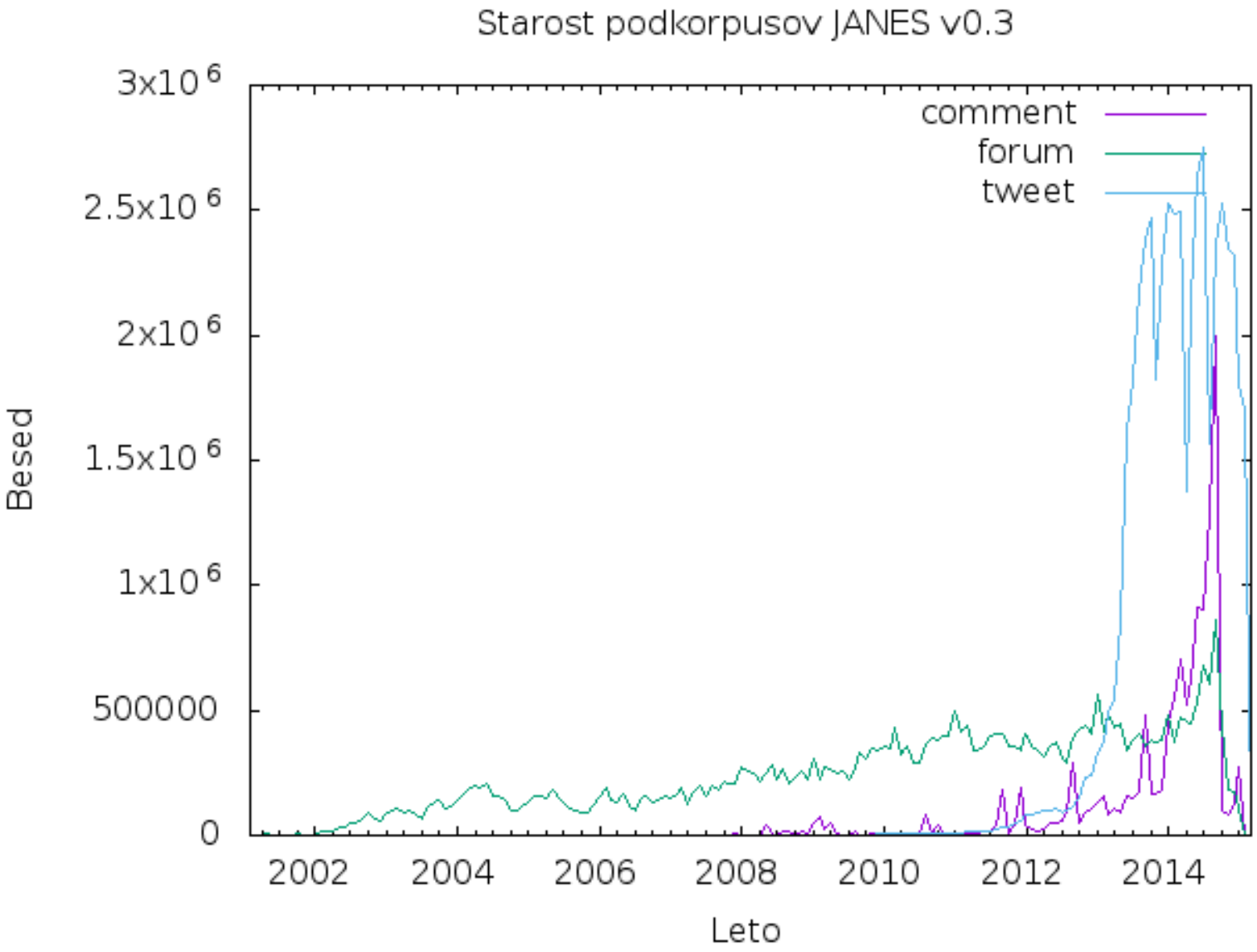

## 135M words




## 161M tokens



# Corpus composition by year



# Corpus processing

- Annotation
  - (almost standard) tokenization & sentence segmentation
  - **lexical normalisation with CSMT**
  - standard MSD tagging & lemmatization
- Encoding
  - (currently) bespoke XML for metadata
  - **annotated text in TEI P5**
- Exploration
  - **(no)Sketch Engine**

# TEI text

```
<s>
    <w lemma="pa" ana="#Cc">pa</w>
    <c> </c>
    <w lemma="še" ana="#Q">še</w>
    <c> </c>
    <choice>
        <orig>
            <w>tamali</w>
        </orig>
        <reg>
            <w lemma="ta" ana="#Pd-fsn">ta</w>
            <c> </c>
            <w lemma="mali" ana="#Agpmpn">mali</w>
        </reg>
    </choice>
    <pc ana="#Z">.</pc>
</s>
```

# noSketchEngine: “better”

Iskalni niz **boljše** **27,257** > Premešaj **27,257** (169.0 na milijon)

Prva | Prejšnja Stran 2 od 1,363 Pojdi Naslednja | Zadnja

| | | | | |
|---|---|---|---|---|
| blog | distancirali tudi terminološko, če se jih ima večina itak za | boljše | /boljše/dober/Agcmpa | od novinarjev in njih cenzorskih, politično nastavljenih |
| tweet | nič spornega http://t.co/hsLiD7A6Dv ##ɠ @BozoPredalic | Boljše | /boljše/dobro/Rgc | za marsikoga, da je ne. Lahko katero prime, da ga |
| blog | tangu. Sicer je super, sedaj me zanima, če je lahko še | boljše | /boljše/dober/Agcfpn | . Grrrr ɠ heh, skrajni čas, bejbi.jst to že nekej časa |
| tweet | Affleck bo naslednji Batman. Buuu ##ɠ A ne bi bilo ful | boljše | /boljše/dober/Agcnsn | , če bi tvite z deli, zaposlitvami opremili z enim |
| forum | rdečic po telesu, včasih tudi po obrazu. Sedaj so vse | boljše | /boljše/dober/Agcfpn | , niso več rdeče le srbi jo še večkrat (včasih se |
| tweet | @maticslapsak Vsakemu svoje veselje. Kaj č'mo. Vseeno | boljše | /boljše/dobro/Rgc | kot nazi ikonografija. #alwayslookonthebrightsideoflife |
| tweet | slovenske oblasti in sodišča ji stojijo na poti v | boljše | /boljše/dober/Agcnsa | življenje http://t.co/6LE4s7GEzb ##ɠ Vsaka tretja ženska |
| forum | malo neprijetno. Savine so sicer v snegu bile veliko | boljše | /boljše/dober/Agcfpn | , na mokri in mastni podlagi pa prava katastrofa v |
| blog | 13.09.2012 ob 16:57 ɠ ne vem, meni se zdi, da bi veliko | boljše | /boljše/dobro/Rgc | (in bolj seksi) izpadlo, če bi imela zgornji del |
| forum | kaksne narezane diske (npr. ATE, brembo... ) in pa | boljse | /boljše/dober/Agcmpa | zavorne ploscice. ##ɠ Pri nekaterih avtih je res potrebno |
| tweet | koga ##ɠ @Razdelilec tista Gradišnikova je huda, ja. | boljš | /boljše/dobro/Rgc | da nima prav ##ɠ hm, a ni Al Gore 07 pokasiral Nobelove |
| blog | dalje), ker morda v takem moodu kaj lepega zamujaš ... ɠ | boljše | /boljše/dobro/Rgc | da neham besedičit ..... lačna sem pa se hočem zamotit |
| forum | cevi, če bi bilo morda res kej v dovodu nafte, pa nič | bolš | /boljše/dobro/Rgc | ɠ - ni 4motion ɠ - kompresija bi avto skos zajebavala |
| tweet | ##ɠ @TamaraSvetina Sem zelo iz vaje, a če ne najdeš | boljše | /boljše/dober/Agcmpa | (ga) se lahko potrudim. @ales_gantar ##ɠ @_Inja _ Kaj |
| forum | Tudi meni ni všeč Astra enjoy, sport mi pa je. Mnogo | boljše | /boljše/dober/Agcmpa | sedeže ima, pa el. ročno in tudi ni veliko dražji |
| tweet | pol sm si pa kupu fruštk in sm še zmer u minusu... | Boljš | /boljše/dobro/Rgc | da bi šou u ošterijo: D http://t.co/SHzwwSnpKF ##ɠ |
| comment | koncano najmani srednje sole. Prekmurski Romi imajo | boljse | | razmere, vecina romskih naselij ima elektriko, vodo |
| forum | njegovih | | | |
| tweet | za st | | | |
| forum | original | | | |

Prva | Prejšnja Stra

**text.type** forum
**text.author** Goggy
**text.title** VW Passat BKP trese - NUJNO POMOČ
**text.date** 2011-03-15
**text.url** http://www.avtomobilizem.com/forum/viewtopic.php?f=6&t=84268&start=20#p1423667
**text.id** janes.forum.avtomobilizem.6.84268.1423667

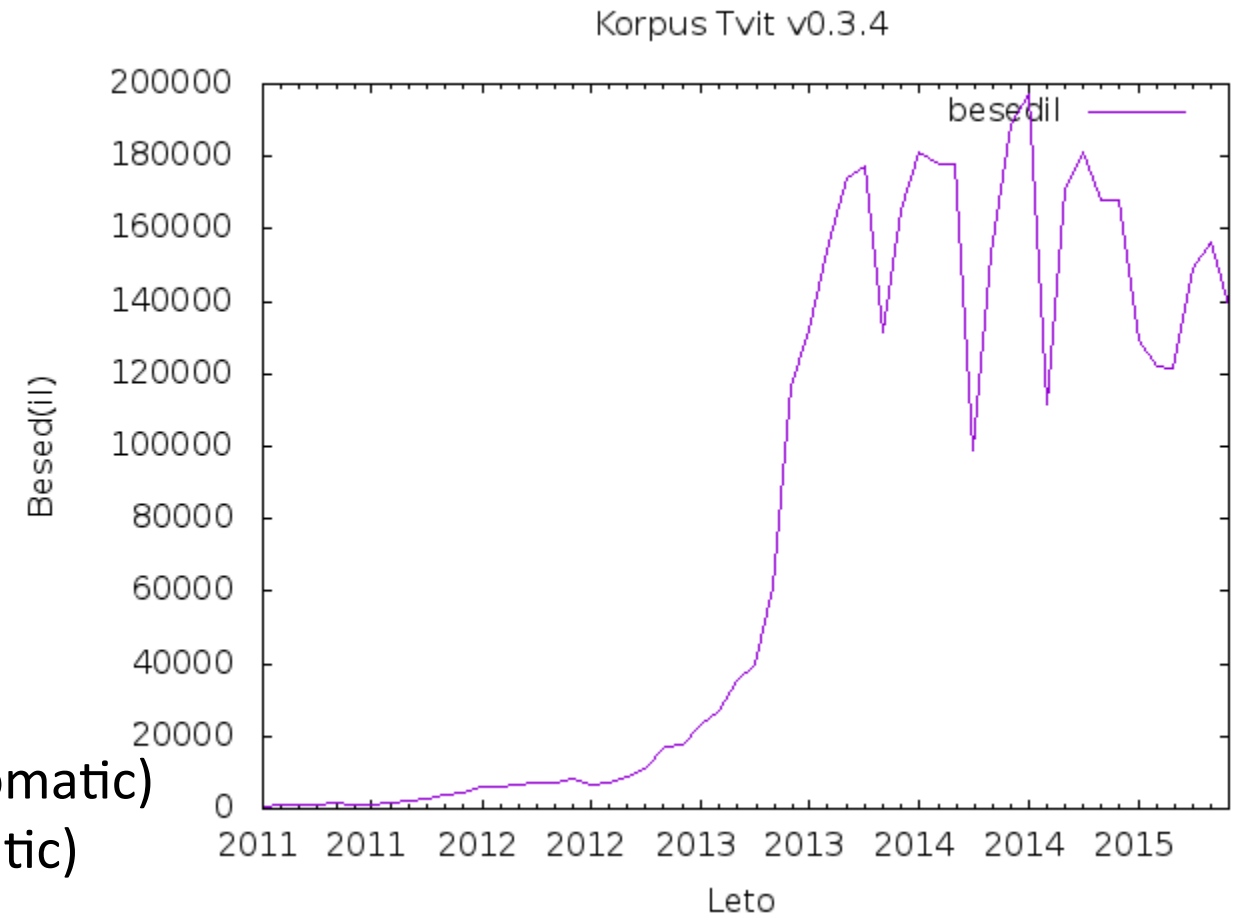
# noSketchEngine: “l”

| | word | Frekvenca |
|---|---|---|
| p \| N | jaz | 224,166 |
| p \| N | jst | 15,654 |
| p \| N | js | 11,256 |
| p \| N | jes | 1,034 |
| p \| N | jez | 308 |
| p \| N | vaz | 208 |
| p \| N | jezt | 62 |
| p \| N | ges | 22 |
| p \| N | jaaz | 10 |
| p \| N | jzt | 9 |
| p \| N | ioz | 8 |
| p \| N | jaaaaz | 7 |
| p \| N | naz | 6 |
| p \| N | jaaaz | 6 |
| p \| N | jiz | 4 |
| p \| N | joz | 2 |
| p \| N | jaaaaaaz | 2 |
| p \| N | iaz | 2 |
| p \| N | iez | 1 |

# 3. Construction & annotation of the subcorpus of tweets

# Corpus Tweet-sl 0.3.4

- Additional tweets
  - tokens: 70M
  - words: 53M
  - tweets: 4,3M
- Enriched metadata
  - at tweet level:
    - standardness (automatic)
    - sentiment (automatic)
  - at user level:
    - private/corporate (manual)
    - male/female (manual)
    - region (automatic)



# Annotation of text standardness (Ljubešić et al. 2015)

- 2 standardness levels
  - technical: T1 – T3
  - linguistic: L1 – L3

**T=1 / L=3**

Original: *Ma men se zdi tole s poimenovanji oz s poslovenjenjem imen mest čist mem.*
Standardised: *Meni se zdi to s poimenovanji oz. s poslovenjenjem imen mest čisto mimo.*

**T=3 / L=1**

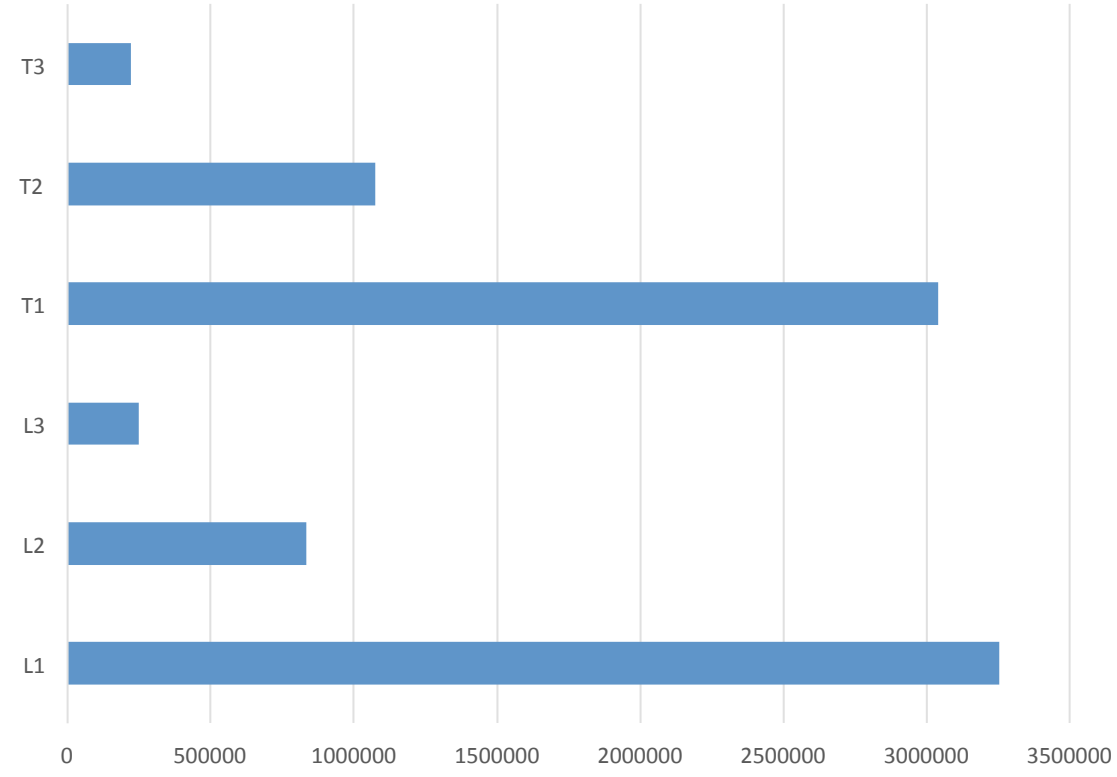
Original: *se pravi,da predvidevaš razveljavitev*
Standardised: *Se pravi, da predvidevaš razveljavitev?*

# Annotation of text standardness (Ljubešić et al. 2015)

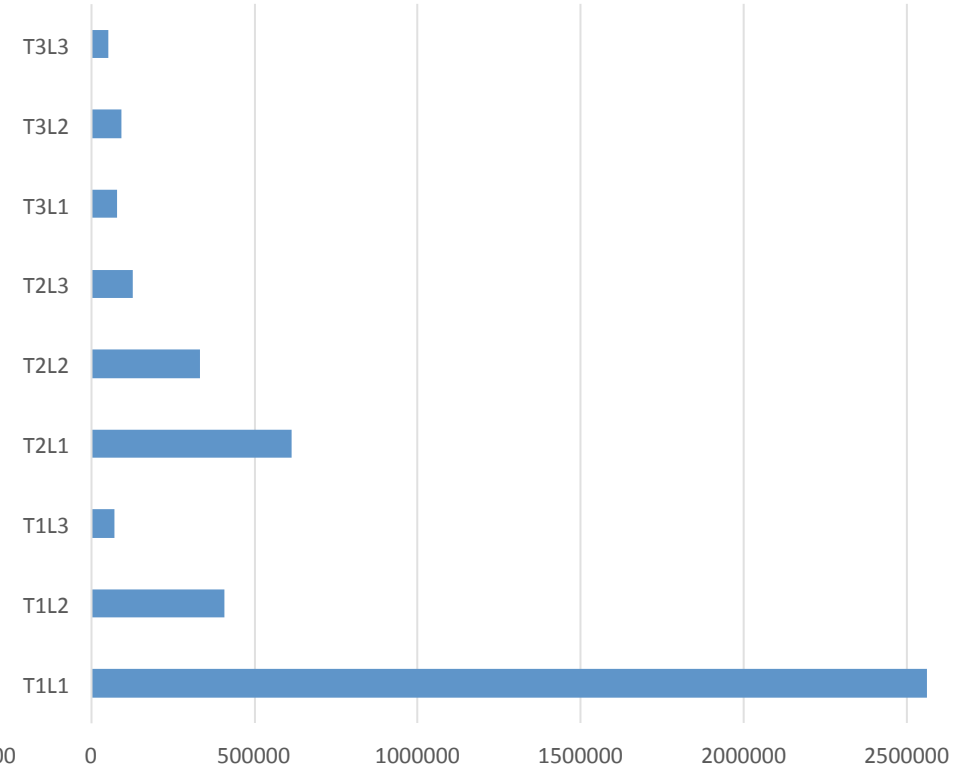
- Dataset development
  - 50:50 standard & non-standard texts
  - ratio of no. of normalised tokens vs. no. of all tokens per text (0.1)
  - manual annotation
    - 900 texts single annotated (development set)
    - 400 texts double-annotated (testing set)
- Feature selection (29 features)
  - character-based:
    - repetitions of characters
    - ratio of alphabetic vs. non-alphabetic characters
    - ratio of vowels vs. consonants
  - token-based:
    - proportion of very short words
    - proportion capitalised words
    - proportion of words not included the Sloleks lexicon
- Regressor
  - grid search hyperparameter optimisation via 10-fold cross-validation on the SVR regressor using RBF kernel
- Evaluation
  - mean absolute error
    - 0.377 T
    - 0.424 L

# Corpus composition by standarndess

## T & L standardness

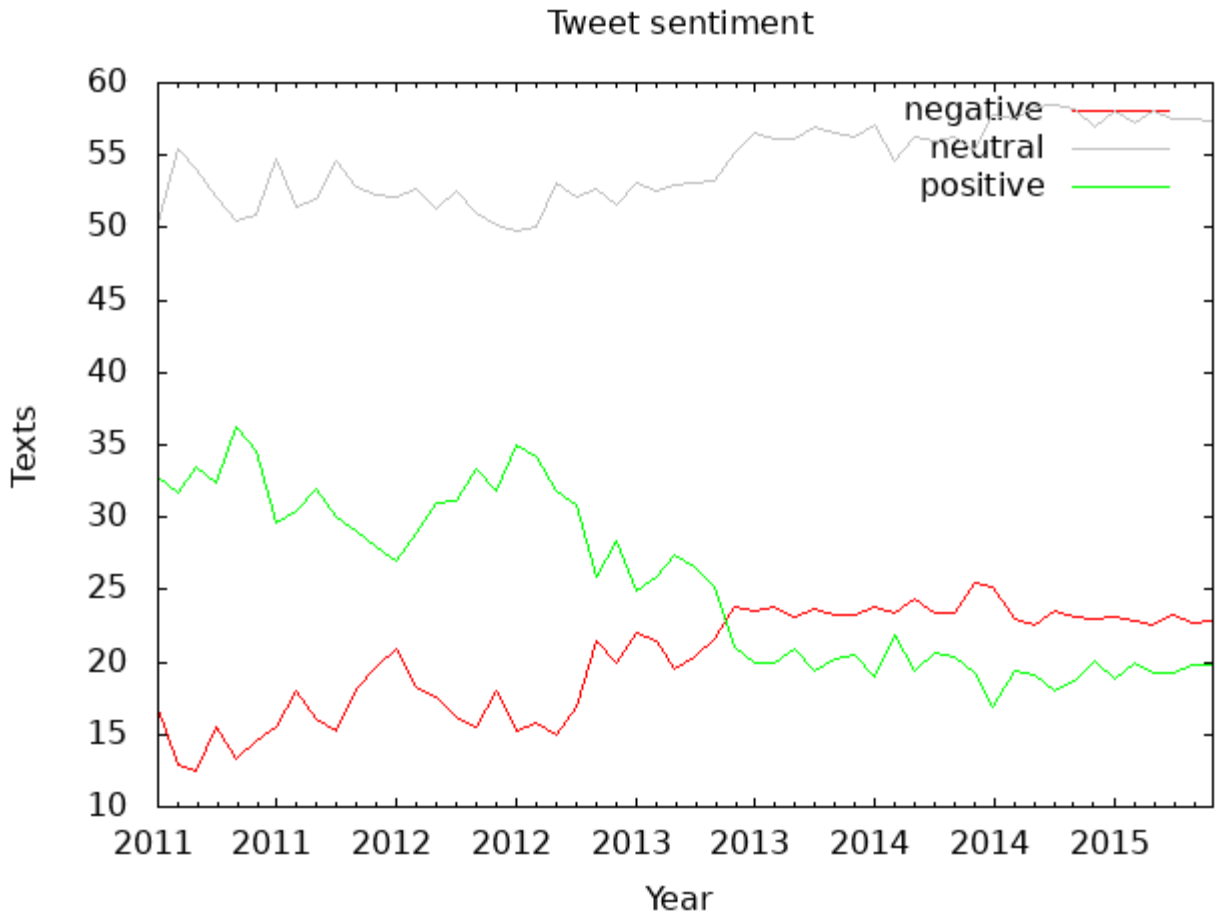



## Combined standardness

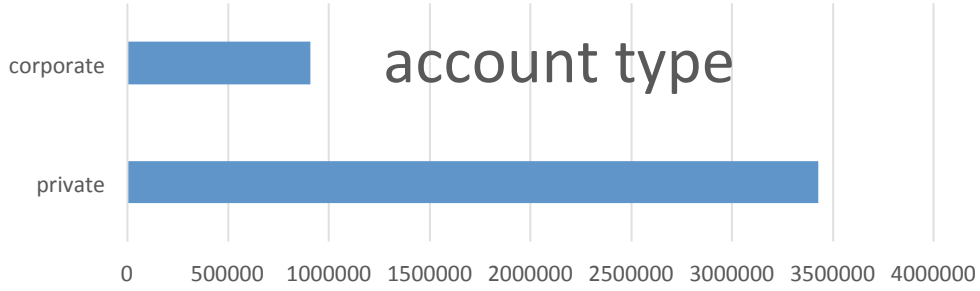


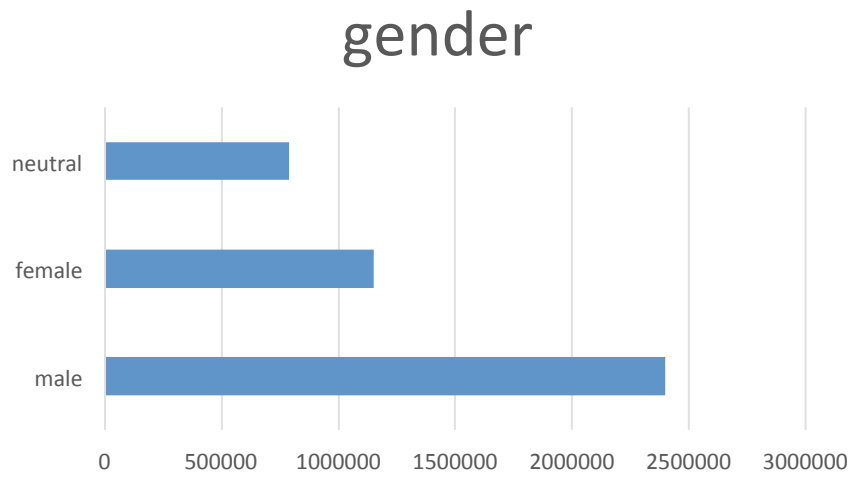
# Annotation of sentiment (Jasmina Smailović)

- Automatic annotation
  - large manually annotated dataset
  - SVM
- Evaluation on 1,000 tweets (sports & politics)
  - Baseline = 37.7%
  - 1-annotator ~ 57.3%
  - 2-annotator = 62.1%
  - IAA = 76.5%



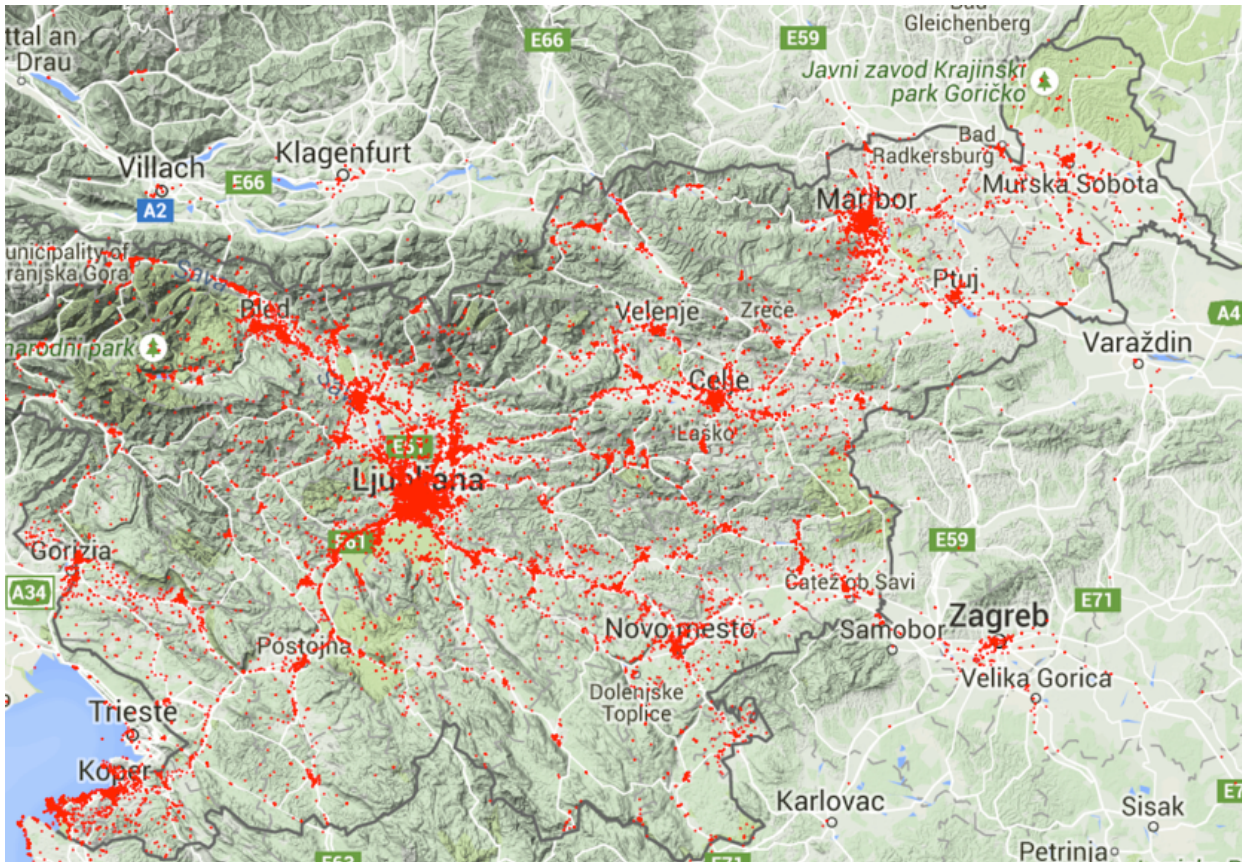
# Corpus composition by account type & gender



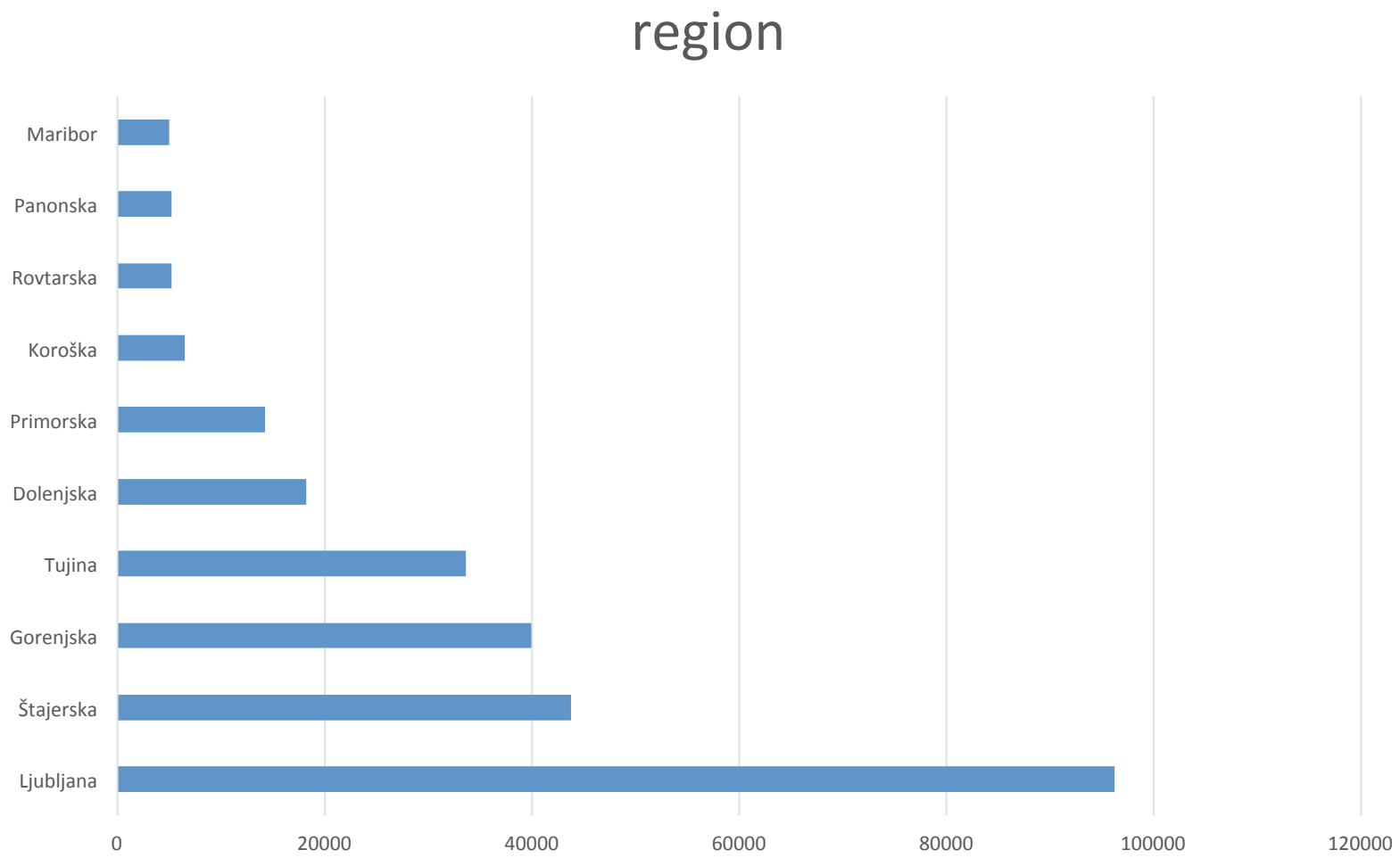



# Annotation of regions (Čibej & Ljubešić 2015)

- Data collection
  - harvesting ny Aug 2015
  - 130K geo-located tweets
  - 1,7K unique users
- Region annotation
  - ray-casting
  - 7 regions + Lj & MB
  - strict filtering
    - private accounts
    - non-standard tweets
    - ≥3 tweets sent
    - ≥90% tweets sent from the same region
  - learning corpus
    - 370 users (6,3%)
    - 75K tokens



# Corpus composition by geographic region



# Tweet-sl 0.3.4 metadata

**id**="tid.392972411765018626"
**name**="007_delic"
**created**="2013-10-23T11:14:58"
**retrieved**="2013-10-24T05:20:32.036451"
**favorited**="0"
**retweeted**="0"
**in_reply**="tid.392965997352591361"
**lang**="sl" lang_prob="0.9967788622457"
**standard_tech**="T1" standard_tech_n="1.1"
**standard_ling**="L1" standard_ling_n="1.2"
**sentiment**="neutral"
**source**="private"
**sex**="female"
**geo**="-"

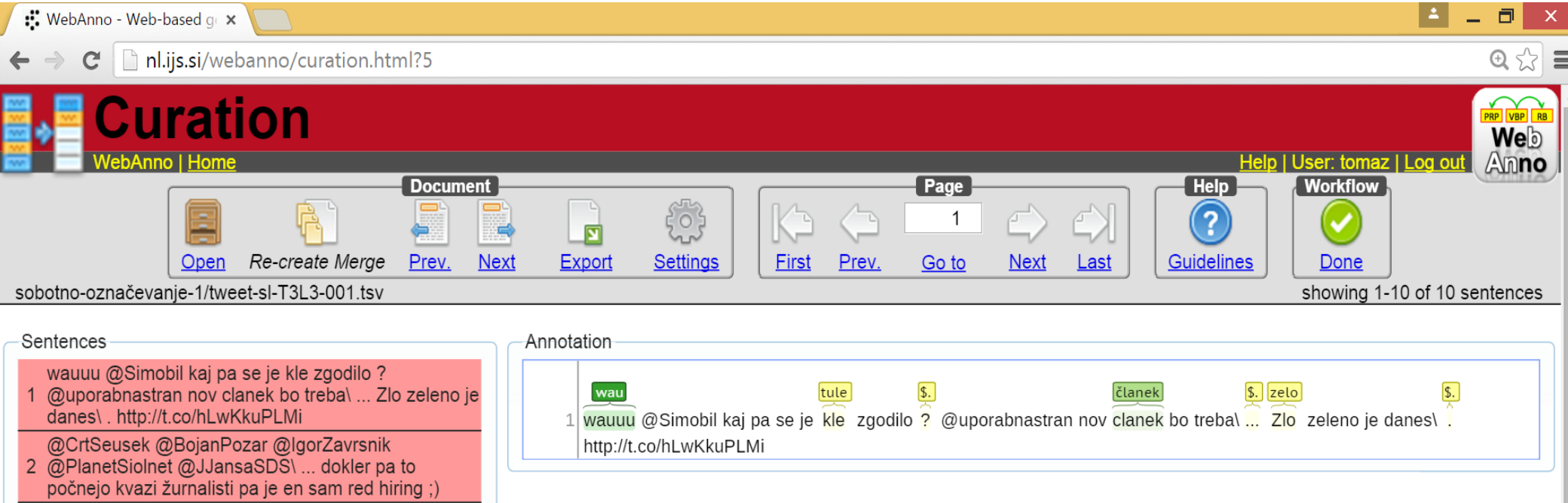
# 4. Work in progress & future work

# Rediacritization (Ljubešić et al., submitted)

- Dataset development
  - Wikipedia, web texts & tweets
  - ≥ 100 characters per sentence
  - marked with L2 & L3 standardness level
  - ≥ 20% tokens with diacritics
- Training set
    - token-aligned parallel dataset with original token & token with stripped diacritics
- Approaches
  - lexicon approach
    - most frequent translation in training data
  - corpus approach
    - translation & language model combined
- Results
  - baseline: 88%
  - charlifter 93%
  - lexicon approach: 98.2%
  - corpus approach: 99,1%
- Error analysis
  - ambiguities, tokenization errors, proper names
  - contextual (syntactic) information needed

# Manual annotation of the Janes reference corpus

- Goal
  1. **tokenization & sentence splitting**
  2. **standardization**
  3. POS-tagging & lemmatization
- WebAnno
  - Annotation guidelines
  - Annotator training
  - Curation



# To-dos - Linguistics



- Lexicology
  - annotate foreign & adopted words
  - compile a dictionary of twitterese
  - detect semantic shifts & neologisms
- Sociolinguistics
  - profanity
  - offensive language
  - flaming

# To-dos - NLP

- Corpus development
  - add Wikipedia discussion & user pages
  - add blogs
  - create subcorpora (e.g. politicians, celebrities)
  - develop a monitor corpus
- Corpus annotation
  - improve normalization, tagging & lemmatization
  - CMC-aware tagset extension
  - add metadata (e.g. age)
- Corpus dissemination
  - promise
    - searchable through an on-line concordancer
    - downloadable as a dataset
  - problems
    - terms of use issues
    - copyright issues
    - privacy issues
  - solutions
    - annonimysation, shufling, sampling
    - different access levels

# 5. Project activities

# JANES Summer Camp

- Uni Lj, 24-28 Aug, 2015
- 25 high school students from all over Slovenia
- Format
  - 5 days, 5 topics
  - lecture, exercise, project
  - project presentation
- Invited talks and evening events
- On-line slides and other teaching materials
- Good media coverage

# JANES Conference

- Uni Lj, 25-27 Nov 2015
- 15 reviewed papers, 23 authors, 50 delegates
- Events
  - Best student paper award
  - Invited lecturer
    (Michael Beißwenger, TU Dortmund)
  - Panel discussion
    (What is Janes Slovene?)
  - Tutorial on statistics and R for linguists
    (Maja Miličević, Uni Belgrade)

# JANES Express

- Stops:
  - Zagreb, Croatia (4 Dec 2015)
  - Belgrade, Serbia (10 Dec 2015)
  - in cooperation with ReLDI
- 1-day event:
  - student workshop (noSkE)
  - annotation workshop (WebAnno)
  - evening lecture (annotating UGC corpora)
  - 150 participants

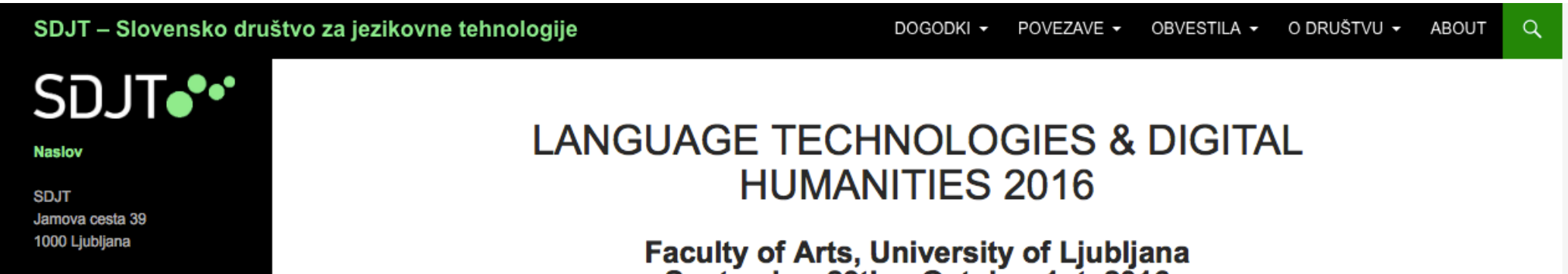
# Journal Special Issue 2016

- Topic: computer-mediated communication
  - construction & distribution of CMC corpora
  - tools & resources for processing of CMC
  - corpus analyses of CMC
  - comparisons of CMC with standard and/or spoken discourse
  - sociolinguistic studies of CMC
  - code-switching in CMC
  - neologism & semantic shift detection in CMC
  - offensive language in CMC
- Deadline: 31 March 2016
  - scientific, survey & position papers
  - reviews & project reports
- CFP: click here

Slovenščina 2.0

# Int. CMC conference 2016

- Topic: computer-mediated communication
  - development of CMC corpora
  - annotation & analysis of CMC corpora
  - NLP for CMC corpora
- Deadline:
  - extended abstract: 1 March 2016
  - full paper: 1 June2016rts
- CFP: click here



# http://nl.ijs.si/janes/

tenks ☺